SGD: The Role of Implicit Regularization, Batch-size and Multiple-epochs

Cornell University

Satyen Kale* Ayush Sekhari† Karthik Sridharan†

{* -Google Research, NY, † - Cornell University }

Google AI

Motivation

SGD is the method of choice for training large scale and over-parameterized machine learning models:

- Small batch size.
- Multiple passes over the dataset.

Why does multi-pass, small batch-size SGD work so well in practice?

Preliminaries

Stochastic optimization:

- Model parameters $w \in \mathbb{R}^d$, data instance $z \in \mathcal{Z}$.
- Given samples $S = \{z_1, \ldots, z_n\}$ drawn i.i.d. from some unknown distribution $\mathcal{D}$, find the point

$$\widehat{w} \in \arg\min_w F(w),$$

where the objective $F(w) := \mathbb{E}[f(w; z)]$.

Stochastic Convex Optimization (SCO): Simplest stochastic optimization problem.

Assumptions:

1. Population loss F is convex and L-Lipschitz.
2. Initial point distance to optimality: $\|w_1 - w^\star\| \le B$ where $w^\star \in \arg\min_w F(w)$.
3. Bounded gradient variance: $\sup_w \mathbb{E}_{z \sim D} \|\nabla f(w, z) - \nabla F(w)\|^2 \le \sigma^2$.

Algorithms:

Stochastic Gradient Descent (SGD): For $i = 1$ to n:

$$w_{i+1}^{\mathrm{SGD}} \leftarrow w_i^{\mathrm{SGD}} - \eta \nabla f(w_i^{\mathrm{SGD}}; z_i).$$

Return $\widehat{w}^{\mathrm{SGD}} := \frac{1}{n}\sum_{i=1}^n w_i^{\mathrm{SGD}}$.

SGD upper bound for SCO (Nemirovski and Yudin 1983)

On any SCO problem, running SGD algorithm for n steps with step size $\eta = 1/\sqrt{n}$ has the rate

$$\mathbb{E}_S[F(\widehat{w}_n^{\mathrm{SGD}})] - \inf_{w \in \mathbb{R}^d} F(w) \le O\Big(\frac{1}{\sqrt{n}}\Big).$$

Regularized Empirical Risk Minimmization (RERM): Given a regularization function $R(w)$ and the empirical loss $\widehat{F}_n(w) := \frac{1}{n}\sum_{i=1}^n f(w; z_i)$, return

$$w_{\mathrm{RERM}} = \arg\min_{w \in \mathcal{W}} \widehat{F}_n(w) + R(w).$$

Main Contributions

1. For any regularizer, we provide a SCO problem for which **RERM fails to learn**. On the other hand, SGD learns at rate of $O\Big(\frac{1}{\sqrt{n}}\Big)$.
2. **Statistical separation** between SGD (small batch size) and learning via Gradient Descent (GD) on empirical loss (large batch size).
3. Provide a **multi-epoch variant of SGD** commonly used in practice.
   (a) This algorithm is at least as good as single pass SGD.
   (b) But, can be much better for certain SCO problems.

SGD, RERM and implicit regularization

SGD and implicit regularization: Is SGD biased towards good solutions??

Conjecture: There exists a regularization function $R(w)$ such that:

$$\widehat{w}^{\mathrm{SGD}} \approx w_{\mathrm{RERM}} := \arg\min_{w \in \mathcal{W}} \widehat{F}_n(w) + R(w).$$

Many positive results: Gunasekar et al. (2018), Soudry et al. (2018), Ji and Telgarsky (2018), Arora et al. (2019) ...

Our result: RERM fails to learn for SCO

For any regularizer R, there exists a SCO problem for which

$$\mathbb{E}_S[F(w_{\mathrm{RERM}})] - \inf_{w \in \mathbb{R}^d} F(w) \ge \Omega(1),$$

where w_{RERM} is RERM solution with regularization $R(w)$.

SGD has no implicit regularization for SCO - the simplest stochastic optimization setting!

Loss function: For $x \in \{0,1\}^d$, $y \in \{-1, 1\}$ and $\alpha \in \{e_1, \ldots, e_d\}$,

$$f(w; z = (x, \alpha, y)) = y\|(w - \alpha) \odot x\|.$$

Data Distribution: $x \sim \mathrm{Uniform}(\{0,1\}^d)$, $y = +1$ w.p. $3/4$ and $y = -1$ w.p. $1/4$, $\alpha = e_1$.

Key Idea: $F(w)$ is convex, but $\widehat{F}_n(w)$ is not convex.

1. There exists a coordinate $\widehat{j} \in [d]$ such that $x[\widehat{j}] = 0$ for all samples where $y = +1$.
2. Along the coordinate $\widehat{j}$ the empirical loss looks like:

$$\widehat{F}_n(te_{\widehat{j}}) \approx -\frac{1}{3}|t| \qquad \text{(non-convex function)}$$

3. **Empirical loss decreases as we increase t**
   ⇒ ERM diverges to infinity.
   ⇒ RERM fails for any regularizer $R(w)$.
   ⇒ Gradient descent algorithm eventually fails (diverges along the direction $\widehat{j}$)

Gradient Descent (Large-batch) vs SGD (Small-batch)

SGD with large batch size / GD:

For $t = 1$ to T:

$$w_{t+1}^{\mathrm{GD}} \leftarrow w_t^{\mathrm{GD}} - \eta \nabla \widehat{F}_n(w_t^{\mathrm{GD}}).$$

Return $\widehat{w}_T^{\mathrm{GD}} := \frac{1}{T}\sum_{i=1}^T w_t^{\mathrm{GD}}$.

Sample complexity lower bound for GD algorithm

There exists a SCO problem such that for any choice of step size and iterations,

$$\mathbb{E}_S[F(\widehat{w}_T^{\mathrm{GD}})] - \inf_{w \in \mathbb{R}^d} F(w) \ge \Omega\Big(\frac{1}{n^{5/12}}\Big).$$

On the other hand, SGD with step size $O(\frac{1}{\sqrt{n}})$ has a rate of $1/\sqrt{n}$.

Proof based on novel modifications of the iteration complexity lower bound construction in Amir. et. al. 2021 to rule out GD small step size.

Single pass SGD vs Multiple pass SGD

Our multi-pass SGD algorithm: Run k passes of SGD + cross validation

Multiple passes can help!

Let $R(\cdot)$ be any regularization function, there exists a SCO problem for which:

(a) **(One pass SGD lower bound)** For any step size η, SGD has lower bound

$$\mathbb{E}_S[F(\widehat{w}_n^{\mathrm{SGD}})] - \inf_{w \in \mathbb{R}^d} F(w) \ge \Omega\Big(\frac{1}{\sqrt{n}}\Big).$$

Furthermore, SGD with $\eta = 1/\sqrt{n}$ attains this bound.

(b) **(Benefit of multiple-pass SGD)** Multi-pass SGD algorithm with k satisfies:

$$\mathbb{E}_S[F(\widehat{w}^{\mathrm{MP}})] - \inf_{w \in \mathbb{R}^d} F(w) \le O\Big(\frac{1}{\sqrt{nk}}\Big).$$

(c) **(Failure of RERM)** RERM algorithm has lower bound of $\Omega(1)$.

Connections to deep learning

Consider a two layer diagonal neural network:

$$h(w; x) = \mathrm{ReLU}(w_2^\top \mathrm{ReLU}(w_1 \odot x)),$$

and absolute loss function $f(w; z) = |y - h(w; z)|$ (also holds for linear loss / hinge loss). There exists a data distribution over (x, y) such that:

- SGD succeeds in finding an approximate global minima.
- There exists a bad ERM solution.